DOCUMENT RESUME

ED 279 693 TM 870 080

AUTHOR Reckase, Mark D.
TITLE Position Paper on the Potential Use of Computerized Testing Procedures for the National Assessment of Educational Progress.
PUB DATE 4 Sep 86
NOTE 20p.; One of 46 papers commissioned by the Study Group on the National Assessment of Student Achievement and cited in Appendix B to their final report "The Nation's Report Card" (TM 870 049). For other papers in this group, see TM 870 050-094.
PUB TYPE Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; *Adaptive Testing; *Computer Assisted Testing; Disabilities; *Educational Assessment; Educational Testing; Elementary Secondary Education; National Surveys; Test Construction; Testing Problems; *Testing Programs
IDENTIFIERS *National Assessment of Educational Progress

ABSTRACT

The current technology of computerized testing is discussed, and a few comments are made on how such technology might be used for assessing school-related skills as part of the National Assessment of Educational progress (NAEP). The critical feature of computerized assessment procedures is that the test items are presented in interactive fashion, allowing the examinee and the computer to alternate in transmitting information. Two of the more popular of the many possible procedures are computerized adaptive testing (CAT) and computerized personality assessment. Advantages of CAT (and other computerized assessment procedures) include flexibility in item selection and administration time, efficiency, greater test security, and clerical processing power. Disadvantages include the cost of the computer equipment, amount of needed computer storage power, and the quality of graphic presentations on the cathode ray tube screen. Other factors relating to computerized testing are: (1) item types; (2) dimensionality of tests; (3) sampling of the content domain; (4) effects of the interaction of mode of presentation and test item; (5) equating of procedures, especially CAT, with less precise paper and pencil tests; (6) test quality--balancing test length versus precision; (7) item pool characteristics; (8) item selection; (9) test scoring; (10) determining the final item (test length); (11) human factors; and (12) the impossibility of omitted items. Testing of students with disabilities is a promising application of computerized assessment for NAEP. (GDC)

Position Paper

on the Potential Use of Computerized Testing Procedures for the National Assessment of Educational Progress

Mark D. Reckase

American College Testing Service

Paper commissioned by

THE STUDY GROUP ON THE NATIONAL ASSESSMENT OF STUDENT ACHIEVEMENT

1986

Position Paper
on the Potential Use of Computerized Testing Procedures for the National Assessment of Educational Progress

Mark D. Reckase

ACT

The purpose of this paper is to review the current computerized testing technology specifically with regard to how it might be used for assessing school related skills as part of the National Assessment of Educational Progress (NAEP). Consideration will also be given to how this technology might change in the future in ways that are relevant to NAEP. In order to develop a framework for the concepts presented in the paper, several components of the current NAEP program will be discussed, followed by detailed comments concerning computerized assessment procedures.

The NAEP was designed to serve as a national evaluation of elementary and secondary education in the United States. A logical consequence of its purpose is that it predominantly focuses on the assessment of school learning. Thus the assessment devices are measures of achievement and are related to very specific outcomes of the educational process. These factors have serious consequences for the assessment process because they imply that fairly complex skills are being assessed rather than pure measures of what have traditionally been labeled as "aptitudes."

The focus on the total elementary and secondary curriculum also has the consequence that the number of skills being assessed are so great in number that it is literally impossible to measure all of the skills for each individual. This fact has not been a serious problem because the program does not pretend to evaluate each individual, but rather it emphasizes the evaluation of the student population as a whole. Each student is assessed on only a sample of the skills and the individual results are compiled to yield an assessment of the complete domain of skills. Fairly sophisticated administration designs (balanced incomplete blocks) and analysis procedures are used to allow the compilation of national statistics from the separate, relatively limited, individual assessments. It is expected that such a procedure will continue to be required in the future and that computerized assessment devices will have to be able to operate under the same constraints.

The current NAEP program has expanded the scope of the assessment instruments to include more requests for information about demographic characteristics of the student population and noncognitive variables. While the dominant focus of this paper will be on the computerized assessment of achievement, the use of computers to obtain other types of information from students will also be considered.

To some extent, this paper reflects the current NAEP program and its many explicit and implied assumptions. It is certainly possible that the structure and function of the NAEP program will change in the next few years. To the extent that it does, the material presented in this paper may not be directly applicable. However, the general concepts should still prove useful in considering how computerized assessment techniques can be used to assess the academic skills of the nation's youths.

General Characteristics of Computerized Assessment Procedures

Computerized assessment procedures have two general characteristics. First, they present the assessment items in a computer controlled mode. In most cases the presentation is on the screen of a cathode-ray tube as on a computer terminal, a television screen, or a personal computer. However, the presentation can also be by voice-synthesizer, computer controlled slide or micro-fiche projector, or printing terminal. The critical feature is that the items are presented by a computer in interactive fashion--the examinee and computer alternate in transmitting information to each other.

The second general characteristic of computerized assessment procedures is that the computational power of the computer is used directly in the assessment process. The computer can play as minor a role as page turner and response tallier, or it can take full control over the testing process including item generation, item selection, formula scoring of the test, test length determination, report generation, and record keeping. It is in systems of the latter type that the full capabilities of the computer are being used.

These two general characteristics define a very broad class of procedures. This has been done purposely to emphasize that there is no single procedure that can be called the computerized assessment procedure. A great variety of procedures exist and many more are possible. This paper will try to consider the broad class of possible procedures while still providing some detail on the more popular current applications--computerized adaptive testing (CAT) and computerized personality assessment (CPA).

The Advantages of Computerized Assessment Procedures

Computerized assessment procedures have been touted by their proponents because of several advantages. In this section of this paper the advantages that are most likely to be realized in NAEP testing will be described. They include flexibility, efficiency, security, and clerical processing power.

Flexibility

Computerized assessment procedures typically store test items in computer memory for selection and administration. Current computer storage media can store large numbers of items in a very easily accessible form. Laser disk technology, in particular, allows the computer access to extremely large numbers of items. The storage capacity allows almost an unlimited number of forms to be available for computer administration. Under the CAT administration model, tests can even be constructed during the process of the test administration. Since NAEP requires many forms of assessment devices to be administered, this flexibility of form administration can help overcome the massive printing and distribution requirements. The computer can administer any test form that can be constructed from the items in its item pool. Since the full component of the NAEP items can be stored on a laser disk, any specific form can be constructed and administered by any computer with access to the disk. If the NAEP tests could all be administered on computer, the printing of multiple forms would be unnecessary. Any form could be administered on any computer. The complete test battery would be available for administration at any location.

A second component of the flexibility of computerized assessment devices is the individual, self paced nature of the assessments. Since the tests are administered "one-on-one" by the computer, the administration to groups is not necessary. The test can be individually scheduled for each student. The tests also do not need to have the same time limit because the computer keeps track of each student's performance and records their test responses. Booklets do not have to be distributed or collected. Of course, this type of flexibility can only be availed if the examination situation allows it. The NAEP environment may not allow the use of flexible scheduling for most students, but it may be helpful for special cases such as adult examinees and handicapped students.

Efficiency

Computerized assessment procedures have the potential for acquiring more information about a student's capabilities per unit time and are therefore be more efficient. This potential has been well documented using the CAT methodology. CAT has been shown to give equivalent measurement precision to standard paper-and-pencil tests in half the time and using on the average one-sixth as many items, or the procedure can supply higher precision in the same amount of time as traditional tests. These efficiencies can be achieved because the CAT procedure attempts to give each person only the test items that are most appropriate and most informative about his/her particular skill level. By not administering items that are much too easy or too difficult for a person, the amount of information provided per item is much greater than the comparable figure from a traditional paper-and-pencil test, which because of its group testing orientation must include items that cover a broad range of skill levels. Of course, to achieve this level of measurement efficiency

requires some strong assumptions about the interaction between a person and an item. An important question about the NAEP testing is whether the testing environment is consistent with these strong assumptions. This issue will be addressed later in this paper.

Security

One threat to maintaining test security is the existence of a paper-and-pencil copy of the test. By administering the test on the screen of a CRT, this threat is removed. The use of the computer also allows the convenient rearrangement of items to make many test forms, thereby minimizing the opportunity for cheating. Of course, computer storage is not totally secure. However, item pools can be encrypted if necessary and numerous devices are being developed to increase the security of information stored on the computer.

Clerical Processing Power

Many of the features of current large scale testing programs are a result of the clerical requirements of scoring large numbers of tests and recording the results. The multiple choice item was invented to make scoring more efficient, as was the scannable answer sheet. The direct entry of responses into a computer terminal or personal computer can further improve the efficiency of the processing of information. When tests are administered by computer, test forms no longer need to be printed and shipped. Since the responses to the test items go directly into computer storage, answer sheets and scanners are no longer needed. Further, information that could not easily be obtained on a paper-and-pencil test, such as response latency and time-of-

day of the response, can now easily be obtained. All of this information goes directly into the computer for analysis.

The computational power of the computer may result in further advances by making new types of tests possible. CAT procedures are examples of these new types of tests. Their existence was made possible by the ability of the computer to select and score items as they are being taken. Adaptive tests were very cumbersome before adequate computer resources were available. New item types that simulate tasks or that allow free responses are being developed as computer software becomes more sophisticated. The computer may finally free us from the multiple-choice format by making substantial amounts of computer processing power available for use during the process of testing.

The Disadvantages of Computerized Testing Procedures

The most prominent disadvantage of a computerized testing procedure is the need for computer hardware for presenting the test items and processing the results. At the very least, sufficient numbers of computer terminals are needed to administer the tests. The numbers may be very large if large groups are tested at the same time. Smaller number of terminals may be sufficient if testing can be scheduled over a period of time.

Computerized testing procedures also require substantial processing power and computer storage. The current generation of personal computers can easily handle the required tasks if properly configured. However, they may not be able to perform the calibration of items for use with item response theory based procedures. The statistical procedures required for item calibration are iterative procedures that require substantial amounts of computation and

computer storage. Given the rate of change in the computer industry, the next generation of personal computers will likely be able to perform the necessary computation.

A more subtle disadvantage of administering tests by computer is the quality of the presentation of the test materials on the CRT screen. For many types of CRT monitors, the quality of graphic materials (graphs, drawings, pictures, etc.) is much poorer than the corresponding images on paper. Large blocks of print also seem more difficult to read on a screen than on a more traditional medium. These factors can affect the difficulty of the test materials. Future CRT screens may improve in quality, reducing this problem, but for the near future, the quality of the presentation of material on a computer screen will probably be less than that for paper-and-pencil tests.

Other Factors Related to Computerized Assessment

While the major advantages and disadvantages for computerized assessment devices are presented above, there are many other factors that should be considered before using computers for educational assessment. This section of the paper will attempt to give a fairly complete list of these factors.

Item Types

Current computerized assessment procedures tend to use the same types of items as the group administered tests (i.e. multiple choice). This does not have to be the case. Several procedures currently exist that try to use the computer to generate tests that are realistic simulations of real tasks. For example, Dave Vale of Assessment Systems Corp. described a test based on the

simulation of the interaction of an order taker and a client at the 1986 meeting of the American Psychological Association. Such simulation exercises are very popular in the medical community. Interest in these approaches leads to the prediction that the availability of computerized assessment devices will stimulate the development of many new item types and a reduction in the use of multiple-choice tests.

Dimensionality of Tests

There is nothing inherent in the computer administration of tests that places any limitation on dimensionality of the construct being measured. However, if item response models are used as the basis of an item selection or scoring algorithm, as in CAT, than the dimensionality of the test is likely to be an issue. Most IRT models assume that the construct being measured is unidimensional. This assumption allows estimates of ability or achievement to be obtained on a single scale. Many aptitude measures may come sufficiently close to meeting this assumption that it does present a problem. Achievement measures, such as those used in the NAEP, are typically not judged as being unidimensional and they may violate the IRT assumption. Solutions to this problem include dividing a test up into unidimensional sets or using one of several multidimensional IRT models. Neither of these approaches are well developed, but research is being done to produce appropriate methodology for dealing with multidimensional tests. For NAEP testing, the more prudent approach would be to sort the test items into unidimensional item sets, since the methodology for this approach is better developed, rather than attempt to use a multidimensional model.

Sampling of the Content Domain

Three approaches are currently being used to ensure that a test administered on a computer is measuring the appropriate construct. The first is to administer all of the items indicated by the test specifications. In this approach the computer is essentially a page turner. The second approach is to randomly sample items from a well specified domain of content. This approach allows the estimation of the proportion of the total domain of items that can be answered correctly. It does not allow anything about the characteristics of the domain (e.g. dimensionality) other than that it is well defined. This procedure does not require that as many items be administered as the full specification approach.

The third approach is to select items from the pool of possible items to provide maximum information or minimum standard error. This approach depends on IRT models and requires their assumptions. This approach is very efficient in the use of items, resulting tests may be very short, but there has been some concern that the selection procedure "purifies" the construct being measured. That is, the construct being measured is forced to be unidimensional even when it is not. Achievement items selected by this approach may not do a good job of representing the entire domain of content.

Mode Effects

Items that are administered on a computer do not always operate the same as the same item administered in paper-and-pencil form. Some items are easier on computer, others are easier in paper-and-pencil form. Some personality researchers have indicated that responses to computer presented inventories tend to be more extreme. Collectively, the interaction of mode of presentation and test item is called a mode effect. If such effects are

present, they make the use of norms or item calibration results obtained fr
a different mode questionable. Scores from a computer presented NAEP test
not be comparable to those from a paper-and-pencil version if a mode effect
exists.

Equating

Computerized testing procedures, particularly CAT procedures, have the capability of controlling the characteristics of test scores to a higher degree than paper-and-pencil tests. For example CAT procedures can be programmed to administer test items to a person until their ability estimate has a prespecified standard error of measurement. Thus, all individuals, no matter what their level of ability, would be measured to the same level of accuracy. It is virtually impossible to achieve this same level of precisio for all persons using a paper-and-pencil test because of the fixed nature of the test.

In order to equate two tests, they should measure the same construct an have the same error of measurement for all persons. Otherwise, examinees wi poor levels of skills should take the less precise test and those with high skill levels should take the more precise test to maximize their scores. Th implication is that tests with unequal reliabilities cannot be equated. The best that can be done is that scores can be made comparable at several point along the score scale.

This theoretical result, which has been proven by Frederic Lord, is the basis for a dilemma. If the full advantage is made of a CAT procedure and measurement is improved, the resulting scores cannot be equated to the corresponding paper-and-pencil test. However, the CAT procedure could be programmed to match the standard error structure of the paper-and-pencil

test. In that case, the scores can be equated, but full advantage is not made of the computerized testing technology. Some difficult decisions need to be made regarding the goals of computerized testing and the desirability of equating CAT and paper-and-pencil procedures.

Test Quality

CAT procedures can be programmed to focus either on minimizing testing time and number of items while maintaining precision of measurement, or maximizing precision of measurement in a fixed time period or number of items. These two goals cannot both be accomplished at the same time. For the purposes of the NAEP it is likely that many short tests at a specified minimal level of precision will be the goal.

Item Pool Characteristics

Computerized assessment procedures select items from an item pool stored on a computer accessible device and administer than on some other computer controlled device. If the computer is just a page turner, the item pool is the same as the set of items contained on the traditional paper-and-pencil test. If some other strategy is used to select items, item pool characteristics become more important.

If items are selected randomly from a well defined domain, it is important that the items represent that domain, or the computerized test will not be a valid measure of the domain. For CAT procedures, the items should not only represent the test specifications, but they should also be evenly spread over the range of ability over which it is desirable to measure accurately. Research by Wayne Patience and Mark Reckase indicates that 200 items, evenly spaced in difficulty, are needed to accurately measure over a

six standard deviation range (i.e., from -3 to +3 on a z-score scale). The degree of accuracy they used in deriving this number was the standard error of a test with reliability, .90. This standard error was required at each point on the sacle between +3 and -3.

Item Selection

CAT procedures cannot be expected to precisely represent an item pool unless they are specifically programmed to do so. If items are merely selected to maximize precision, the selection process may miss important concepts. The item selection program should be written to sample from each specification component to insure that all content areas are covered.

Two different item selection algorithms are currently popular with CAT procedures. The first is to select items that provide the most statistical information at the current ability estimate. The second is to select the item that minimizes the posterior error variance of an ability estimate in a Bayesian estimation procedure. The former procedure seems to be gaining favor among the individuals applying CAT.

Test Scoring

Computerized assessment devices have the major advantage of scoring the test immediately. If the computer is simply presenting items in the same form as the paper-and-pencil test or by sampling items from a domain, this scoring is done exactly as it is for the paper-and-pencil test. Of course, the computer can also immediately convert the score to a standard score or percentile, but there is nothing new in the process. If a CAT procedure is used, scoring is usually done using a maximum likelihood or Bayesian estimation procedure. The Bayesian procedures have the advantage of giving an

updated estimate after every item is administered, but they also tend to regress estimates because a Bayesian prior is used. Maximum likelihood estimation requires that both a correct and an incorrect response be present before an estimate can be obtained, but the procedure does not use a prior sc regression effects are not present. The characteristics of these estimators should be considered when selecting a procedure.

Stopping Rules

Procedures based on random sampling of items and CAT allow the computer to determine the length of the test. In both cases the tests can either be c fixed length, but with different items for each person, or of variable length with the length determined by the precision required by the testing application. The test lengths can be quite different for different individuals, particularly if the test is only being used to make a pass-fail decision. Those examinees with abilities close to the decision point will have longer tests than those far from the decision point. Also, high ability examinees will tend to have shorter tests than low ability examinees because of guessing effects.

Human Factors

Taking a test on a computer is quite different than taking a test in paper-and-pencil form. The examinee responds using some sort of computer input device rather than by making marks on an answer sheet. There are many ways of entering information into a computer including a keyboard, mouse, joy stick, touch-sensation screen, or special response panel. Each of these methods may change the test slightly. Unfortunately little is known about th effects of each of these response modes on performance. The research does

indicate, however, that a wide variety of types of examinees have little trouble responding to tests on computers.

Another human factor problem is the readability of material on the CRT screen. Large blocks of text seem to be harder to read on a CRT than in printed form. Graphics are also harder to decipher. These factors may differ for each type of CRT, changing the difficulty of the test. Human factor considerations need to be taken account of when designing a computerized assessment system.

Omits

When examinees omit responses on a paper-and-pencil test, they are usually scored as incorrect and no further notice is taken of them. In CAT, the previous response is used in selecting the next item. If an examinee omits an item, no recommended procedure exists for updating the ability estimate and selecting the next item. For this reason many CAT procedures require a response to each item. Omits are not allowed.

This section of this paper has attempted to summarize briefly the issues that need to be considered when using computerized assessment procedures. Little advise has been given about how each of these issues should be resolved because the resolution depends on the particular application. They have been included here to emphasize their importance.

Computerized Assessment in the Future

Several trends in computer technology are likely to have an impact on the use of computers for testing purposes in the future. These are: (a) the

reduction in cost, (b) the increase in computer power in personal computers, (c) the increase in storage capacity, (d) the development of laser disc technology, (e) the development of voice recognition and synthesizer device, and (f) the improvement of graphics capabilities. All of these trends will tend to make computers more available and useful for computerized assessment. Personal computers are already capable of storing and administering moderate numbers of items, some of them with graphics. In the near future, personal computers can be expected to have sufficient storage capacity for any reasonable item pool. Laser disk technology will allow item to be stored as images. Extremely large item pools can be stored in computer accessible form in this manner.

Because of the increases in computing power, more computational intensiv items can be administered by computer. These include items with animation an realistic simulations of real time processes. Multidimensional adaptive test will also be possible resulting in estimates of ability on a number of dimensions at the same time.

Telecommunications technology is also likely to improve. As a result, results of computerized assessments will likely be transmitted directly to a central processing facility. It may also be possible to transmit items to a number of sites by satellite based systems.

One very promising use of computerized assessment is to assist those individuals with some type of disability. A voice synthesizer and computer controlled Braille writer can make the testing of blind students more reasonable. Special keyboards can assist students with motor impairments. Untimed or self paced tests may be useful for students with learning disabilities. The area of special testing is a very promising one for computerized assessment.

The Use of Computerized Testing for NAEP Testing

Computerized assessment is not yet a practical alternative for the evaluation of a large number of examinees because of the need for computer hardware. As schools acquire more personal computers, it may be possible to schedule reasonable size groups at one time. If the computers or CRT's are different at different sites, however, the effect of the type of machine will have to be considered in interpreting the scores.

Computerized assessment procedures are best at sampling a domain, minimizing testing time, maximizing test precision, or assisting disabled students. To the extent that these features are needed in NAEP testing, computerized assessment should be seriously considered. However, since NAEP measures achievement, the issues of dimensionality will have to be considered if IRT based procedures are used. The IRT procedures are somewhat robust to minor violations of the unidimensionality assumptions, so they are likely to be usable--but the reasonableness of the assumption should still be considered.

Testing of students with disabilities seems like the most promising application of computerized assessment in the NAEP. This testing is already done on an individually scheduled basis and with relatively small numbers of students. The use of computer assistance in this application would seem to have important advantages.

This paper has been written as a general overview rather than a scholarly review of the literature. Therefore, references have not been included to the relatively extensive research that exists. To get further detailed information, the following references are suggested.

Green, B. F.; Bock, R. D.; Humphreys, L. G.; Linn, R. L. & Reckase, M. D. (1982). Evaluation plan for the computerized adaptive aptitude battery. Baltimore, MD: Psychology Department, The Johns Hopkins University.

Reckase, M. D. (1977). Procedures for computerized testing. Behavior Research Methods & Instrumentation, 9(2), 148-152.

Wildemuth, B. M. (1984). Microcomputer-assisted testing: resources from ERIC. Educational Measurement: Issues and Practice, 3(2), 48-49.